CHAPTER 14
VALIDITY

The *Standards for Educational and Psychological Testing*[1] describes validity as "the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores." The *Standards* also describes three broad, traditional categories of validity evidence necessary to support such inferences: content-related, criterion-related and construct-related. In the years since the *Standards*, these traditional notions of validity have been supplemented with specific criteria for performance assessments (e.g., Frederiksen & Collins[2], Linn, Baker & Dunbar[3]) as well as the idea that the consequences of using a given test are an important aspect of validity. Consequential validity addresses the issue of whether a test measures what it is intended to measure, or more broadly, whether it has the effect it is intended to have. The description and uses of consequential validity were proposed and advanced by Samuel Messick[4] of Educational Testing Service.

The importance of the consequential validity of the test becomes obvious when one thinks of the alternatives. KIRIS existed because the Kentucky legislature believed that KIRIS would increase the likelihood of KERA's success by supporting (indeed, driving) changes in the classrooms of Kentucky. Thus, all decisions related to KIRIS ultimately had to be considered in light of the question, "Will this change lead to better instruction and more real achievement?" If not, the change became less justified.

INTENDED GOALS OF KENTUCKY ASSESSMENT PROGRAM

The role of the Kentucky assessment program, including the KIRIS assessments, was to promote educational improvement for all children in the state. It did this in three major ways:

1. The assessment program provided goals, standards, and criteria for educational achievement. These goals, standards, and criteria were found linked together throughout the assessment program. They included the statements of goals in the KERA legislation, the academic expectations, and the links between academic expectations and specific items and their scoring guidelines. The assessment program also included operational definitions of success, various performance levels, and relative weights of assessment components.

1American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1985). Standards for Educational and Psychological Testing. Washington: APA. Note that a revised edition of the Standards is in preparation.

2Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. Educational Researcher, 18(9), 27-32.

3Linn, R. L., Baker, E. L. & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. Educational Researcher, 20(8), 15-21.

4Messick, S. (1989). "Validity", in R.L. Linn (ed.), Educational Measurement, Third Edition, MacMillan Publishing Co., 1989.

2. The assessment program provided useful information on progress towards achieving those goals made by schools. Although the major informational use of KIRIS scores was in relation to school accountability as mandated in the KERA legislation, much assessment information was also provided that was useful for monitoring achievement and progress of individual students, the state as a whole, and various subgroups within the state.

3. The assessment program provided useful information on potential differential impact of the assessment program within the school, district, region, and state for various subgroups such as gender, ethnic and racial minorities, and children receiving Title 1 assistance.

EVIDENCE AND INTERPRETATION OF CONSEQUENTIAL VALIDITY

The KERA legislation and subsequent programs, including the assessment program, have engendered much discussion and activity. However, relatively little formal research is available on the impact of the assessment program on classroom practice, teacher development, or support of educational reform in Kentucky. In addition, in the available research it is often difficult to separate effects of the assessment program from other aspects of educational reform. For example, while the criteria for mathematics portfolios made an impact on classroom instruction as a result of being included in the assessment system, some changes resulted from the fact that Kentucky adopted standards that effectively reflected the national standards promulgated by the National Council of Teachers of Mathematics. The understood positive impact of the mathematics portfolio was not sufficient to keep it in the accountability system when it was perceived to be too demanding of time.

It should be noted that the discussion in this chapter of the consequential validity of the Kentucky assessment program paints broad strokes, which may not apply to every classroom in every school. Change was taking place at different speeds and in different forms throughout the Commonwealth, often for different combinations of reasons. In addition, the assessment program, and people's understanding of it, has changed over time. Continuing research will be required to provide more complete results of the consequential validity of the Kentucky assessment program, as well as to keep research findings up to date.

CONSEQUENCES: PROVIDES GOALS, STANDARDS, AND CRITERIA FOR INSTRUCTION AND CURRICULUM. Evidence continues to accrue regarding the effects of KIRIS on instructional practice, teachers' professional development, and support for educational reform. Evidence presented in the *KIRIS Accountability Cycle I Technical Manual* addressed impact on instructional practice, professional development, and educational reform, citing several studies.[5] Taken as a whole, those early studies suggested that KIRIS had an impact on instructional practice.

5Appalachian Educational Laboratory. (Dec., 1994). Instruction and assessment in accountable and non-accountable grades, Notes from The Field, 4(1), 1-2.; Pankratz, R., Ochs, D. et al. (April, 1995). Configuration maps: Results from Kentucky. Papers presented at the annual meeting of the American Educational Research Association, San Francisco, CA; Policy Studies Associates, Inc. (1994). Third-year evaluation of the nine-site program initiative. (A report to the U.S. Department of Education.) Washington, DC:

After the completion of the *KIRIS Accountability Cycle I Technical Manual*, several other studies took place that addressed consequential validity in the context of curriculum and instruction. One major study by RAND[6] was cited in the *KIRIS Accountability Cycle 2 Technical Manual.* This survey and interview-based study found that "about 40 percent of the teachers reported that the open-response items and portfolios have had a great deal of positive effect" on instruction, with only half as many teachers endorsing this view of performance events, and almost none considering multiple-choice items to have had such a positive effect. It should be noted that the study examined a broad range of perceived effects, including perceived negative impact on instruction, and provides a more extensive discussion of its findings than can be afforded in this chapter.

CONSEQUENCES: PROVIDES INFORMATION ON STATUS AND PROGRESS. KIRIS provided information to schools and districts in several forms. These are described more fully in the chapter on Reporting in this *Technical Report*. Essentially, KIRIS reports included scores on each subject matter area and a non-cognitive index for schools and districts each year. KDE and its contractors produced biennial reports that summarized each school's performance in terms of a two year baseline and two subsequent years of the accountability cycle, and summarized the school's status in relation to rewards and assistance. These four year cycles of baseline and accountability overlapped, with each two year accountability block serving as the baseline for the next accountability cycle two years later. Student reports were produced each year for subject matter areas, and the writing portfolio. The student reports were sent by the contractor to schools, where they were distributed to parents by different systems determined by the school. In addition, each year KDE and its contractors produced a summary report for the state, by region, by gender, by ethnic group, and disabilities.

Schools, districts, and classroom teachers reported using the score reports in a variety of ways consistent with the intent of KIRIS. The most common use was in broad program review to mark progress over a year or two, and to focus resources for instructional program improvement. Analysis of KIRIS scores comprised an essential part of every school's annual Transformation Plan. However, KIRIS scores have been used by schools and teachers for other purposes. There were increasing requests to have KIRIS provide information in addition to the school accountability function it was

---

Author; McCollum, H. et al. (August, 1994). Portfolio assessment in mathematics: Views from the classroom. Annual report. Washington, DC: Policy Studies Associates, Inc.;Roberts, R. & Kay, S. (September, 1993). Kentuckians' expectations of children's learning: The significance for reform. A public report prepared for the Prichard Committee for Academic Excellence and the Partnership of Kentucky School Reform, Lexington, KY: Roberts & Kay, Inc. (Available from the Prichard Committee for Academic Excellence, P.O. Box 1658, Lexington, KY 40592-9980.); Winograd, P., Jones, D., & Perkins, F. (submitted). The politics of alternative assessment: Lessons from Kentucky. (Manuscript obtained from first author. ; Laguarda, K. G., Breckenridge, J. S., Hightower, A.M., & Adelman, N. E. (September, 1994). Assessment programs in the statewide systemic initiatives (SSI) International, primary contractor.) Prepared under contract for the National Science Foundation, SRI International, primary contractor. Washington, DC: Policy Studies Associates, Inc.

6Koretz, D. M., Barron, S.,Mitchell, K. J., & Stecher, B. M. Perceived effects of the Kentucky Instructional Results Information System (KIRIS). Santa Monica, CA: RAND.

originally designed to provide. There have been wide-spread calls for additional information in the other traditional evaluation areas.[7] Some examples are listed below:

1. Individual student achievement status for use on report cards;

2. Individual student comparative status for college admissions;

3. School comparisons (the news media routinely convert reports into "rankings" that facilitate comparisons between schools and districts);

4. Student diagnostic information for monitoring student progress and informing instructional changes by classroom teachers;

5. Instructional program evaluation (e.g., to monitor and improve instructional programs, school curricula, and inform teacher assignments and professional development).

While KIRIS provided results that addressed these areas to some extent, most would require substantial changes in KIRIS design and or operation. Some of these uses were possible; and some were possible but somewhat incompatible with the intended uses of KIRIS results. For example, using student KIRIS results as the sole basis for school report card grades was viewed by KDE as an inappropriate use of KIRIS results. Providing diagnostic information for individual students would require not only a complete revamping of the KIRIS test but also a much more rapid feedback than was possible. It should be remembered that KIRIS is a school accountability assessment and that it was not designed as a student diagnostic instrument. KDE believed that such diagnosis was more appropriately undertaken by classroom teachers using classroom assessments other than or, at least, in addition to, KIRIS assessments, and was more appropriately undertaken earlier in the school year (i.e., much earlier than April which is near the end of the school).

CONSEQUENCES: IS FAIR TO SCHOOLS. It is important that the Kentucky accountability program provide a fair educational goal for all schools. This is especially true regarding the consequences of rewards and assistance based on the assessment results. As noted in the *KIRIS Accountability Cycle 1 Technical Manual*, several factors were examined to explore whether the Kentucky accountability program was fair to schools. The factors included geographical location of school, racial/ethnic composition, economic status of students, initial baseline score, school size, and grade level organization. Based on these analyses for the first and second accountability cycles, the Kentucky assessment program appeared to be fair in that rewards and assistance were distributed across these dimensions without statistically significant unevenness. The exception was grade level, where proportionally more elementary schools received rewards than did middle schools or high schools. This result seemed to be explained by

7For example, see the report done by The Evaluation Center, Western Michigan University, (January, 1995). *An independent evaluation of the Kentucky Instructional Results Information System (KIRIS).*

differences in breadth and complexity of the knowledge and skills presented at the different school levels (Elementary, Middle and High) rather than bias.

PROGRAM-SPECIFIC SCHOOL- LEVEL EFFECTS. Beyond being fair with regard to characteristics of student enrollment, KIRIS should not disadvantage schools participating in programmatic efforts to improve student learning. The effort in which Kentucky schools participated most widely was Title I, a federal program established to serve economically disadvantaged students by providing supplemental funding to schools, based on the poverty level of students in the district and school. Between the 1995 and 1998 school years, about 70% of the public schools in the Commonwealth participated. Kentucky was unusual in the nation: as a high poverty state, Kentucky had over half its schools with a reported 50% or more students qualified for free or reduced price school lunch. Approximately 70% of Kentucky students were served by Title I during Cycle 3.

Table 14-1 indicates the numbers of Title I schools participating in school-wide programs, targeted assistance, and the totals and percentages compared to all schools. Table 14-1 clearly demonstrates the decrease in use of Title I as students progress through the school system. Most of this change was the result of decreasing numbers who received free or reduced price lunches, the primary economic criteria for Title I participation. Many reasons have been proposed for this decrease in participation, increasing family wealth with the passing years, and increasing embarrassment over receiving free lunch as the favorite explanations of the decrease in eligibility for Title I.

TABLE 14-1 A1[1] SCHOOL TITLE I PARTICIPATION					
	School Wide	Targeted Assistance	Total Title I Schools	Total Public Schools	Title I Percentage
GRADE 4					
1995	119	595	714	796	89.7%
1996	157	544	701	792	88.5%
1997	418	276	694	786	88.3%
1998	490	179	669	779	85.9%
GRADE 5					
1997	403	272	675	770	87.7%
1998	476	177	653	766	85.2%
GRADE 7					
1997	124	99	223	343	65.0%
1998	153	66	219	338	64.8%
GRADE 8					
1995	30	252	282	354	79.7%
1996	44	192	236	348	67.8%
1997	119	99	218	339	64.3%
1998	148	66	214	334	64.1%
GRADE 11					
1995	1	82	83	236	35.2%
1996	1	39	40	234	17.1%
1997	13	23	36	233	15.5%
1998	24	16	40	237	16.9%
ALL GRADES COMBINED[2]					
1995	129	843	972	1274	76.3%
1996	173	722	895	1274	70.3%
1997	506	372	878	1272	69.0%
1998	608	248	856	1268	67.5%

[1] A1 through A6 schools are defined on page 5 of chapter 10.
[2]The total is smaller than the sum of the grade levels because of the P-8. P-12, and 7-12 schools.

Over the course of the third accountability cycle, elementary, middle, and high schools that participated in the Title I program achieved relatively greater progress toward their KIRIS improvement goals than did non-Title I schools. This is a favorable finding with regard to consequential validity.

CONSEQUENCES: IS FAIR TO STUDENTS. Given that KIRIS was designed to foster learning for all the children in the Commonwealth, it was important that the assessment and the educational system be fair to all students. One crucial fairness question in any testing program is whether the assessment adversely affects one group of students compared with another. If differential performance exists based on race and/or gender, additional investigation is warranted into the assessment or the educational system. Demographic information describing the gender, racial, and students-with-disabilities subgroups of the population of Kentucky's students and the results of a study by Smith, Neff, and Nemes (1999) were discussed extensively in Chapter 11. The study by Smith, et al. used theta scores from KIRIS 1992-1998 to compare the performance of students by gender and racial subgroups. The main findings of this study, detailed in Chapter 11, were that academic performance differences by gender and race exist among Kentucky students. The overall pattern of difference found the following rank ordering of student scores (highest to lowest): White female, White male, African American female, and African American male. Further, in-depth school-based research will be used to follow-up this study.

As detailed in Chapter 4, Dorans and Schmitt[8] standardized mean difference (DIF/SMD) for differential item functioning, comparing groups of male and female students and African American and White students matched on total common-item scores. Results of DIF studies carried out with the 1997-98 found no differences in item functioning between subgroups.

CONTENT AND CONSTRUCT VALIDITY

Although consequential validity concerns may ultimately prove more important than issues of technical quality, it remains very important to examine KIRIS score validity from a traditional psychometric perspective. Thus, content validity information is reviewed below, and construct-related validity evidence is discussed based on the relationships of KIRIS tests with each other, with scores from other testing programs, and with qualitative criteria for judging school quality. Intermingled with traditional notions of validity in this analysis are more recently proposed criteria for evaluating performance assessments: systematic validity, directness and transparency[9]; and, fairness, transfer, generalizability, cognitive complexity, content quality, content coverage, meaningfulness, cost, efficiency, and consequences[10].

8 Dorans, N., & Schmitt, A.P. (1991). Constructed-response and differential item functioning: A pragmatic approach (ETS research report No. 91–47). Princeton, NJ: Educational Testing Service.

9Frederiksen & Collins.

10Linn, Baker & Dunbar.

CONTENT-RELATED VALIDITY EVIDENCE

Section II of this manual describes how the components of the KIRIS assessment were derived from Kentucky's six Learner Goals and the 57 Academic Expectations, using advisory committees of Kentucky teachers to make those outcomes and expectations operational through test items, and to make choices about what the tests would contain. Many tables in Chapter 3 summarize the academic expectations as well as the distribution of the Core Content measured by the items in the assessment, and there is little to add to the extensive treatment of this material. In short, there is substantive validity-related evidence in the process by which KIRIS assessments were constructed.

While test development information serves as the primary source of content-related validity evidence, examining KIRIS tests in terms of the novel content-relevant criteria noted above provides a potential source of additional evidence. Cognitive complexity, content quality, and content coverage can serve as criteria by which to evaluate performance assessments. Because there exist no established standards for these criteria (as noted by Linn, Baker and Dunbar), any detailed consideration of them probably requires discourse substantiated by expert judgment in the form of task analysis. Although content quality *per se* has not been examined, the presence of teacher and other content area specialist judgment in writing and selecting items for the assessments provides some indication of validity in this regard.

CONSTRUCT-RELATED VALIDITY EVIDENCE

Because the KIRIS testing program assessed student performance in several content areas using a variety of testing methods, it is important to study the pattern of relationships among such content areas and testing methods. One method for studying patterns of relationships to provide evidence supporting the inferences made from test scores is the multi-trait, multi-method matrix (see *KIRIS Accountability Cycle 1Technical Manual*). Another method for studying patterns of relationships among varying types of test or item scores is factor analysis. To provide evidence for the construct validity of KIRIS open-response item scores, factor analysis was performed on data obtained from open-response scores from the 1993 through 1996 KIRIS testing program (see *KIRIS Accountability Cycle 2 Technical Manual* for a detailed description of this analysis).

CONCURRENT VALIDITY-RELATED EVIDENCE

Another measure of validity is how well a test correlates with accepted measures of the same or similar constructs. To the extent that few or no other "primarily performance-based" assessments exist for comparison with KIRIS, options are limited for demonstrating concurrent validity--although many traditional measures have been enhanced to include constructed-response supplements in conjunction with multiple-choice or selected-response items. The best one can do is to compare the performance of students on KIRIS to accepted or "traditional" tests of academic achievement, despite the fact that they will not assess exactly the same construct as KIRIS. Correlations with KIRIS should be moderately high, since much of what is measured by most norm

referenced tests have many common KIRIS content area requirements, however, KIRIS has additional higher order thinking requirements for test items that many norm referenced multiple choice tests do not possess. If the correlations are very high, it would imply that the higher order thinking testing requirements in KIRIS are either missing or highly correlated with traditional forms of assessment. If very high correlations occurred, use of KIRIS would be difficult to justify based on the supposition that it not only measured basic but also higher order thinking, since the more traditional forms of assessment would be measuring the same material but much less expensively and with faster turnaround time.

Concurrent validity can be assessed through correlational study using different units of analyses (1) the student, (2) the school, and (3) the state.

STUDENT-LEVEL RELATIONSHIPS. In addition to providing concurrent validity evidence, a good reason for comparing KIRIS to traditional forms of assessment is that those traditional measures are still in use. Tens of thousands of Kentucky high school students take the American College Test (ACT) each year for college admissions. If KIRIS proved to be uncorrelated with that measure, it would place students, parents, and teachers in the uncomfortable position of having to choose the test on which they would like to focus their attention. Hoffman[11] correlated high school juniors' and seniors' ACT scaled scores and their theta scores on KIRIS. Student scores from the years 1994, 1995, and 1996 comprised the data. The sample of students who took the ACT had higher open-response scores than Kentucky students in general. This difference indicated that the results of this study might be generalized to only the upper portion of the distribution of high school juniors and seniors. The strongest relationships were: KIRIS Reading and ACT English scores, r = .56; KIRIS Reading and ACT Reading, r = .52; KIRIS Reading and ACT Composite, r = .56; KIRIS Math and ACT Math, r = .72; KIRIS Math and ACT Composite, r = .70; KIRIS Science and ACT Science r = .57; KIRIS Science and ACT Composite, r = .62. For this group of high-school students (N=51,967), there were moderate to high, positive, linear correlations between these scores. The relationships were stronger between mathematics scores, however, the author reported no test reliabilities and it may be that the higher correlations between mathematics tests were due to higher reliabilities of the mathematics assessments.

Wise[12] described initial results of efforts to link scores from the Armed Services Vocational Aptitude Battery (ASVAB) and KIRIS for Kentucky high school students for the years 1993 through 1996. The number of students matched by year by grade ranged from 3,567 to 16,314. Total students matched for all years were 64,278. Using data for the years 1994, 1995, and 1996; the student-level score (scaled and theta scores for the respective tests) correlations for reading ranged from r =. 51 to r = .56. The correlations for the KIRIS math scores for these years were considerably higher, the range was r = .63 to r = .73. By contrast, the correlations for science were

11 Hoffman, R. G. (1998) Relationships Among KIRIS Open-Response Assessments, ACT Scores, and Students' Self-Reported High School Grades. Radcliff, KY: Hunan Resources Research Organization.

12 Wise, L. L. (1997) Merging ASVAB and KIRIS On-Demand Scores: Report of Preliminary Results. Ratcliff, KY: Human Resources Research Organization.

somewhat lower, ranging from .42 to .58. The correlations for the 1993 KIRIS assessments with ASVAB were somewhat lower than in the other years compared. The author suggested that this might have been due to somewhat lower reliabilities for KIRIS for that year. Other ASVAB tests, designed to measure nonacademic areas of achievement, for example, Auto and Shop Information, did not match KIRIS subject matter. These positive, linear, moderate relationships indicate that KIRIS measures constructs similar to those measured by the content-matching subscales of ASVAB.

SCHOOL-LEVEL RELATIONSHIPS. Considering that KIRIS scores were used for school accountability, it may be argued that a high degree of relationship between KIRIS and other test scores obtained by schools is even more essential evidence of concurrent validity than correlations among student level scores. Obtaining evidence of this kind is problematic insofar as few other tests are administered to all students in a school, in contrast to KIRIS, which was given to about 99% of all students in most years of the third accountability cycle. (For KIRIS, the remaining students were exempted typically for medical or language reasons, or participated in school accountability through the alternate portfolio program, as noted in Chapter 7.) Of the other tests not given to all students, very few are administered to a representative or even approximately random sample of students at participating schools, further diminishing the meaningfulness of school-level scores.

STATE-LEVEL RELATIONSHIPS. Considering that KIRIS was administered only in Kentucky, there were no other states with which to compare student performance on KIRIS. However, it is possible to compare changes in state-level scores on KIRIS over time with state-level changes over time on other measures. The challenge in making such a comparison is that, relative to the first year of testing, some improvement in KIRIS scores was likely to occur as a result of directing school curricula toward the test and familiarizing students with responding to open-response questions in general. Initial gains from the 1992 baseline were unlikely to generalize to other tests, but were a predictable, initial result of implementing a high-stakes testing program. To the extent that this effect has been observed with multiple choice tests[13] used in a high-stakes setting, a finding that initial KIRIS gains did not generalize to other tests was not evidence against score validity, but rather an indication that caution must used in interpreting score gains relative to the first year of high-stakes testing.

The best available comparison in this regard is the National Assessment of Educational Progress (NAEP). During Cycle 3, Kentucky participated in the NAEP program. NAEP is a standards based assessment that is administered to a national sample. NAEP is also administered at the state level, to a different sample of students. The state assessments are not aggregated to obtain the national results. Kentucky has participated in all of the assessments since NAEP began state testing in 1990. A thorough discussion of the NAEP results compared with KIRIS results can be found in Chapter 2 of this Technical Report. However, it should be noted that there are methodological issues related to scaling in making comparisons across measures. Not

13See, for example, Linn, R. L. (1995). Assessment-based reform: Challenges to educational measurement. Educational Testing Service: Princeton, NJ.

only is each test built to its own specifications, but also each measure has its own scale. As long as each measure provides an indication of whether changes over time are statistically significant, it is possibly to compare trends broadly. Comparing the magnitude of changes on one measure with magnitude of changes on another is more complicated, especially when multiple sets of scores are available for one or the other of the measures (such as theta and standards-based--Novice, Apprentice, Proficient, Distinguished-- scores on KIRIS open-response tests).

CRITERION-RELATED VALIDITY EVIDENCE

A vast potential source of validity evidence to support or refute the inference that accountability gain scores reflect improvements in school performance is schools themselves. The primary challenge associated with taking advantage of this rich source of information is that it is logistically difficult (and therefore expensive) to gather meaningful data on schools. A lesser challenge (and, some would argue, a potential advantage) is that information of this nature does not necessarily lend itself readily to quantification, so that results must be considered mostly in qualitative terms.

A case study of 16 schools conducted by Kelley[14], a senior research associate at the Wisconsin Center for Educational Research, provided important initial evidence for criterion-related validity. Kelley found that successful schools had taken specific actions to achieve success, including analyzing test results to identify weaknesses, setting goals, changing curriculum and using professional development effectively. By contrast, low-success schools had not changed their curriculum and had not used professional development to learn about the new learning goals. These were favorable results for KIRIS, but more data are necessary. Additional evidence will be gathered to study the relationship between KIRIS score gains and changes in school practice.

14Kelly, C. and Protisik, J. (1997). Risk and reward: Perspectives on the implementation of Kentucky's school-based performance award program. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

This page was intentionally left blank.